

UNDERGRADUATE PROJECT PROPOSAL

|  |  |
| --- | --- |
| **Project Title:** | **Toxic Comments Classification Based on Deep Learning** |
| **Surname:** | **Gao** |
| **First Name:** | **Ruiling** |
| **Student Number:** | **201918020301** |
| **Supervisor Name:** | **Joojo Walker** |
| **Module Code:** | **CHC 6096** |
| **Module Name:** | **Project** |
| **Date Submitted:** | **November 4, 2022** |

**Table of Contents**

[1 Introduction 3](#_Toc118902827)

[1.1 Background 3](#_Toc118902828)

[1.2 Aim 4](#_Toc118902829)

[1.3 Objectives 4](#_Toc118902830)

[1.4 Product Overview 4](#_Toc118902831)

[2 Background Review 4](#_Toc118902832)

[2.1 Summary of existing approaches 4](#_Toc118902833)

[2.2 Brief summary of related literature 5](#_Toc118902834)

[3 Methodology 5](#_Toc118902835)

[3.1 Approach 5](#_Toc118902836)

[3.2 Technology 6](#_Toc118902837)

[3.3 Version management plan 6](#_Toc118902838)

[4 Project Management 6](#_Toc118902839)

[4.1 Activities 6](#_Toc118902840)

[4.2 Schedule 7](#_Toc118902841)

[4.3 Data management plan 7](#_Toc118902842)

[4.4 Deliverables 7](#_Toc118902843)

[5 References 8](#_Toc118902844)

# Introduction

## Background

With the advent of the digital age, the number of electronic documents is increasing. One of the more typical phenomena is that with the growth of online communities on the Internet, a large amount of cluttered textual information, such as posts, comments, etc., floods the entire online environment. Most normal and virtuous netizens are spontaneously maintaining the sustainability and usability of the Internet, but in contrast, there are some users who post anti-social and malicious comments on online platforms in an attempt to undermine the usability of the Internet [1]. The emergence of malicious comments violates the legitimate rights of netizens and can cause serious mental and psychological harm to them. Therefore, the efficient management and processing of large volumes of comment texts have become a target of interest for researchers [2].

Text classification is one of the effective ways to locate and triage information efficiently and accurately, solving the problem of information clutter as much as possible [3]. Text classification (TC), also known as text categorization, is an extensive area of current research in linguistic text mining and processing. TC is a process that uses deep learning algorithms to categorize text content into pre-given sets of labels [4]. Deep learning-based text classification techniques have been developed and matured since the 1990s. Compared to text classification systems based on knowledge engineering and expert systems, classification techniques using deep learning provide better classification results and flexibility and have become the main techniques used in related fields [5]. Among the techniques for text classification by different criteria, sentiment analysis (SA), also known as opinion mining, is a branch of text classification. Its main function is to identify and analyze the sentiment in a text by using pre-given labels with human sentiment colors and sentiment tendencies, such as positive, negative, neutral, etc [6].

The report is divided into four sections, and the structure of the report and the main content of each section are organized as follows. The first section introduces the basic concepts of text classification, the purpose and significance of the research, an analysis of the problems addressed by text classification, and an overview of the research on the topic. Finally, the overall structure of the report is given. Section 2 introduces the research background of text classification and summarizes the current state of research and the main features of text classification. Section 3 presents the main research methods chosen for the topic, the techniques used, and the version management plan. The final section presents the research steps and plan of the topic.

## Aim

The main goal of this project is to develop deep learning-based models for the detection and classification of toxic comments automatically.

## Objectives

The objectives of this text classification project are as follows:

* Conduct background research on text classification, understand the field and the corresponding technologies.
* Collect usable dataset from the Internet.
* Clean and pre-process the data for modeling.
* Extracting features from the text in the cleaned datasets.
* Train different models using datasets and assessing the quality of the models
* Analyze the quality of the models and compare the strengths and weaknesses of each model

## Project Overview

The product is designed to analyze the sentiment of comments made by users on social media platforms, filtering out malicious comments and classifying them into different categories, such as hate speech, personal attacks, pornography, or violence. The product helps social network staff to automatically screen out malicious comments, reducing labor and time costs, while also helping to clean up the online environment.

# Background Review

## Summary of Related Literature



Table 1: existing approaches and their feature

The table illustrates the features of existing approaches.

To date, several models have been used in research on sentiment classification. Among them, Wei, G and Wu, K carried out text feature extraction of product reviews by a weighted word vector approach and subsequently modelled the data using support vector machine (SVM). By using weights, the dimensionality of the vectors was controlled to effectively retain the important words in the textual information, and the sparsity problem in traditional models was solved [6].

Tang X, Tang J and Zhou, A used principal component analysis (PCA) for the features retained after completing lasso regression analysis, combined with a Random Forest hybrid model (RF), to detect malicious comments in online Chinese communities. The experimental results show that the model combined with PCA is more accurate and more suitable for malicious comment detection than the traditional random forest algorithm [7].

Fan N, An Y, and Li, H have proposed an analysis method that divides the sentiment of a text into local and global sentiment, determining the local sentiment of the text through a conditional random field model and subsequently analyzing the global sentiment of the text using the K-nearest neighbor algorithm (KNN). The advantages of choosing this method for text sentiment analysis are the multi-level nature of the sentiment analyzed and the high accuracy that can be achieved [1].

# Methodology

## Approach

This report is designed to investigate the effectiveness of different deep learning models for predicting malicious comments by modelling the same dataset and comparing the accuracy of different models. By modelling the same dataset and comparing the accuracy of different models, the most accurate model is selected as the most suitable machine learning model for malicious comment classification. The models are listed as follows.

Support Vector Machine algorithm (SVM) is a machine learning model based on statistical principles. Its main principle is to find a hyperplane in space that is more than sufficient to delimit all data samples and to make the distance from all data in this dataset to this hyperplane the shortest [8].

To address the shortcomings inherent in a single model in machine learning, the Random Forest algorithm (RF) integrates multiple decision trees into a forest, where each decision tree is a classifier, and the random forest will aggregate all the classification votes to predict the outcome [7].

K-Nearest Neighbor algorithm (KNN), a widely known statistical method. He determines the class of a new text by identifying texts that have already been classified and, after a new text has been input, considering the class of the K texts that are closest to the new text from which the trained text was derived [9].

This paper focuses on training and tuning the parameters of these different models, and finally evaluating their performance in terms of accuracy, prediction rate, recall and other evaluation methods. The data used for this study is a 68.5M English comment dataset crawled from social media platforms on the Internet, which contains the ids of the posters, the content of the comments posted, and the different types of toxicity. By using the same dataset with the same data pre-processing progress, controlling variables to ensure the type of model as single variable, so that the performance of different models can be analyzed as accurately as possible.

## Technology

The hardware of the project is MacBook Pro Apple M1 Pro 16 GB.

Python 3.1.8 and Jupyter Notebook are used to implement the methods.

## Version management plan

Whenever a new version of the code or related electronic documentation is updated, it will be uploaded to the Baidu Cloud in order to keep track of all project progress.

# Project Management

## Activities

Tasks to be done to complete objects are listed as follows:

| Objects | Activity |
| --- | --- |
| Ob1: Conduct background research on text classification, understand the field and the corresponding technologies. | A1.1 Identify subject keywords |
| A1.2 Search for relevant essays |
| A1.3 Read the relevant literature |
| A1.4 Summary the advantages and limitations of different technologies |
| A1.5 Perform a literature review |
| Ob2: Collect usable dataset from the Internet. | A2.1 Search for social media comment datasets on Kaggle |
| A2.2 download the datasets |
| A2.3 Identify the structure of the datasets |
| Ob3: Clean and pre-process the data for modeling. | A3.1 search for methods to clean the data |
| A3.2 apply methods on datasets |
| A3.3 evaluate the process |
| Ob4: Extracting features from the text in the cleaned datasets. | A4.1 Search for methods to extract data |
| A4.2 Implement methods on the datasets |
| A4.3 Evaluate the feature extracting methods |
| Ob5: Train different models using datasets and assessing the quality of the models | A5.1 Search for documentation of different models |
| A5.2 Apply models in Python language |
| A5.3 Adjust the parameters until the model performs optimally |
| Ob6: Analyze the quality of the models and compare the strengths and weaknesses of each model | A6.1 Search for different methods to evaluate the model |
| A6.2 Apply multiple rubrics to different models |
| A6.3 Put the results into a table |

Table 2: Activities towards the objects

## Schedule

The schedule that need to be completed throughout the project and their due dates are listed in the chart below:

图表, 条形图

描述已自动生成

Table 3: Schedule of the project period

## Data management plan

Data of the project is planned to be stored in the Baidu Cloud service in the purpose of keeping track of the progress of the project.

## Deliverables

Throughout the execution of this project, the following items will be submitted for assessment:

* Project proposal with ethical forms, showing detailed description of the work to be done.
* Project log containing planned objectives for each week.
* Progress report providing justification of the project.
* Project presentation illustrated by a poster and a practical demonstration.
* Final report which comprises a complete and clear explanation of the problem to be solved.

# References

[1] N. Fan, Y. An, and H. Li, “Research on Analyzing Sentiment of Texts Based on K-nearest Neighbor Algorithm,” *Computer Engineering and Design*, vol. 33, no. 3, pp. 1160–1164, 2012, doi: 10.16208/j.issn1000-7024.2012.03.053.

[2] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, “Feature selection via maximizing global information gain for text classification,” *Knowl Based Syst*, vol. 54, pp. 298–309, 2013, doi: 10.1016/j.knosys.2013.09.019.

[3] Li R, “Research on Text Classification and Its Related Technologies,” Apr. 2005.

[4] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, “Text categorization: past and present,” *Artif Intell Rev*, vol. 54, no. 4, pp. 3007–3054, Apr. 2021, doi: 10.1007/s10462-020-09919-1.

[5] J. S. Su, B. F. Zhang, and X. Xu, “Advances in machine learning based text categorization,” *Ruan Jian Xue Bao/Journal of Software*, vol. 17, no. 9, pp. 1848–1859, Sep. 2006, doi: 10.1360/jos171848.

[6] G. Wei and K. Wu, “Sentiment Analysis Based on Word Vector Model,” vol. 26, no. 3, 2016, doi: 10.15888/j.cnki.csa.005655.

[7] X. Tang, J. Tang, and A. Zhou, “Research on Community Malicious Comment Detection based on a Hybrid Model of Feature Selection and Random Forest,” *Morden Computer*, vol. 2020, no. 07, pp. 22–26, 2020, doi: 10.3969/j. issn. 1007-1423. 2020. 19. 005.

[8] J. Cui, J. Liu, and Z. Liao, “Research of Text Categorization Based on Support Vector Machine,” *Computer Simulation*, vol. 30, no. 2, pp. 299–302, 2013.

[9] N. Zhang, Z. Jia, and Z. Shi, “Text Categorization with KNN Algorithm,” *Computer Engineering*, vol. 31, no. 8, pp. 171–185, 2005, [Online]. Available: http://www.xteamlinux.com.cn,